DOCUMENT RESUME

ED 194 598 TM 800 729

AUTHOR Travers, Eugene J.
TITLE Some Uses of Standardized Test Results. Director's Handbook Topics in Testing Measurement and Evaluation.
INSTITUTION New Jersey State Dept. of Education, Trenton. Div. of Operations, Research, and Evaluation.
PUB DATE 79
NOTE 14p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Aptitude Tests; Criterion Referenced Tests; Educational Diagnosis; Elementary Secondary Education; *Grouping (Instructional Purposes); Norm Referenced Tests; *Predictive Measurement; *Standardized Tests; *Test Interpretation; Test Results; *Use Studies

ABSTRACT

The uses of standardized test results are discussed and various measurement terms are defined. Scores on criterion referenced tests are interpreted by relating performance to some predetermined criterion, whereas scores on norm-referenced tests are interpreted by comparing them with scores obtained by members of the group on which the test was normed. Standardized test results may be used for: (1) diagnosing learning difficulties; (2) detecting discrepancies between potential and achievement, e.g. scholastic aptitude tests (which are general in nature and measure one's ability to improve as a result of instruction) and achievement tests (which are specific in nature and tend to be used to evaluate an individual's status upon the completion of training); (3) assessing growth in academic achievement; (4) grouping students for instruction according to some attribute that may be measured (in conjunction with other sources of data such as cummulative records, report card grades, and teacher judgement); (5) planning instructional activities; and (6) determining reasonable achievement levels in order to modify the curriculum. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194598

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

DIRECTOR'S HANDBOOK
TOPICS IN TESTING MEASUREMENT AND EVALUATION

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
L. F. Price
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800 729 NEW JERSEY STATE DEPARTMENT OF EDUCATION
DIVISION OF OPERATIONS, RESEARCH AND EVALUATION
Fred G. Burke
Commissioner of Education

SOME USES OF STANDARDIZED TEST RESULTS

Eugene J. Travers

Introduction

In the following discussion of uses of standardized test results, various measurement terms are included which may require definition and clarification for some readers. In general these terms are briefly defined and clarified as they are introduced in the text. However, a few of them may require further discussion beyond that included in the main body of the text. They are discussed here in the introduction.

Item analysis: This term usually refers to a process used by test makers to evaluate single test items by various methods. These usually include determining the difficulty value and discrimination power of the item, and possibly its correlation with some criterion. As used in this section of the Handbook, however, it refers to a commonly available score report which indicates the number and percent of students who responded correctly on each test item for a specified group of students (class, school, school district, etc.). Sometimes it also indicates the number and percent of students who chose each possible wrong response and the number and percent who made no response to the item.

Criterion-referenced tests: Both norm-referenced and criterion-referenced tests measuring a given curriculum area must sample from the same universe of possible test items for that area. However, criterion-referenced tests are likely to deal with a smaller content area than are norm-referenced tests and to have more items measuring that limited area. Scores for groups or individuals on criterion-referenced tests are interpreted by relating performance to some predetermined criterion such as a specified proportion of items correct. Because they tend to have many items measuring a limited area, criterion-referenced tests are often more suitable for diagnosing pupils' strengths and weaknesses in that limited area than are comparable norm-referenced tests.

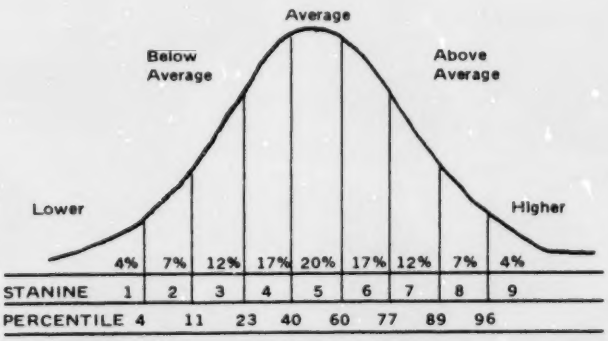
Norm-referenced tests: Norm-referenced tests sample a broader, more comprehensive area of content and contain fewer items for measuring each of the various components than criterion-referenced tests do. Scores of individuals and groups are interpreted by comparing them with scores obtained by members of the group on which the test was normed. The norm group usually consists of scores from a national sample, or the scores obtained in the local school district that year. For example, it might be indicated that a student's score was at or above the level of 60% of those in the norm group. In the case of national norms, this would mean that the student scored at or above the level of 60% of

students in the national sample on which the test was normed. In the case of local norms it would mean that the student scored at or above the level of 60% of the students in the local school district who took that particular test that year.

Percentile rank: A percentile rank tells what percent of pupils in the norm group scored the same or lower than a given score on the test. If a given score has a percentile rank of 60, then 60 percent of the pupils in the norm group scored the same or lower than that score. A percentile rank is **not** a "percent correct." It is not the percent of questions the pupil answered correctly but rather the percent of pupils in the norm group who scored at or below that score. Equal differences in percentile ranks do not always represent equal differences in the characteristic being measured, therefore these scores cannot be added, subtracted, multiplied or divided. This severely limits their use for research purposes. However, these scores are clear and simple to understand and provide an excellent means for reporting pupil performance to parents.

Stanine scores: A stanine is a score on a nine-unit scale from 1 to 9, where a score of 5 represents average performance. The highest stanine is 9; the lowest is 1. Stanines represent equal amounts of the characteristic being measured, except for stanine 1 and 9. Stanine 8 is as far above average as stanine 2 is below average. Figure 1 that follows shows the percentage of students that tend to fall in each stanine in a normal distribution. It also shows the relationship between percentiles and stanines.

FIGURE 1
Normal Curve



Since each stanine represents a range of raw scores, students, teachers, and parents are not likley to attribute unwarranted significance to small raw score differences. This makes these scores extremely useful for reporting pupil performance at parent-teacher conferences. Stanines 4, 5, and 6 represent average performance. Stanines 7 and 8 represent above average performance and stanine 9 represents performance which is far above average. Stanines 2 and 3 represent below average performance and Stanine 1 represents performance far below average. Stanines like other norms, compare a pupil's performance with that of other pupils. They do not describe performance in terms of any fixed criterion.

Grade equivalent scores: A grade equivalent score indicates the grade level in years and months, for which a given raw score was the average score in the standardization sample. For example, a raw score of 20 with a grade equivalent of 4.5 means that, in the norm group, 20 was the average score for pupils in the fifth month of the fourth grade. However, the test publisher may not actually have tested any norm-group students during the fifth month of the fourth grade. The grade equivalent may actually be an estimate based on testing done in some other month. After the test has been standardized, if another pupil in the fifth month of the fourth grade were to obtain a raw score of 20, his/her performance would be judged at grade level or average for his/her grade placement. If he/she were to obtain a raw score of 25 or a grade equivalent of 5.1, he/she would be performing like a typical student in the first month of the fifth grade **on that same test.** This does not mean that the fourth grader is now ready to do all fifth grade work. There are many things a fifth grader has done that are not measured on the fourth grade test. In a similar manner a grade equivalent score of 3.5 for the fourth grader would indicate that he/she was performing like a typical student in the fifth month of the third grade **on that same fourth grade test.** It does not indicate that he/she has learned only third grade material.

Grade equivalent scores are popular with many educators because they **seem** clear and easy to understand. However, they lend themselves to a variety of misinterpretations. It is recommended, therefore that they **not be used** to interpret student performance on standardized tests. Other types of scores such as percentile ranks and stanines are more suitable for that purpose.

Diagnosing Learning Difficulties

Standardized test results can be quite useful for diagnosing learning difficulties both for individuals and for groups. This is true whether the test be norm-referenced or criterion-referenced. A norm-referenced test compares a student's performance with that of other students on the same test, while a criterion-referenced test compares a student's performance with a predetermined standard. Examples of standardized achievement tests available from the major publishers which are useful for this purpose are the Stanford Achievement Test and the Iowa Test of Basic Skills.

Let us first concern ourselves with diagnosing group difficulties. Standardized test results are reported in various forms. The type of report that is perhaps most useful for diagnosing group learning difficulties is the item analysis. The term "item analysis" often refers to the test maker's process of evaluating single test items by a variety of methods. As used here, however, it refers to a common type of score report available from the publishers which indicates for each item the percentage of students in a specific group (class, school, district, etc.) who responded correctly. Sometimes, it also indicates the percentage of students who chose each possible response to the item. As a further aid to interpretation some item analyses have their items organized in clusters which measure a similar skill or specific segment of knowledge.

For example: On a mathematics test, all of the items which measure multiplication of whole numbers might be grouped together. By looking at the item analysis for that cluster of items one can immediately get an indication of how well the group has mastered that specific skill. Sometimes the average percentage of correct responses on the items within each cluster is reported.

By examining the item analysis for a specific group of students, whether it be a classroom, a school or a school district, one can very quickly determine where the strengths and weaknesses lie, which skills the students seem to have mastered very well, and which skills seem to cause them difficulty. This can be quite useful in planning learning strategies for the future. In reference to the example cited, if one finds that students have mastered most of the whole number operations on a mathematics test, but show rather marked weaknesses in certain aspects of division of whole numbers, this may be an indication that some change in instructional methods or materials is needed, or simply that teachers need to concentrate more heavily and provide more encouragement on that particular topic. If the item analysis also indicates the number and percentages of students who chose each of the wrong responses, clues may be provided concerning the specific misunderstandings students have in the particular area of difficulty. It is often useful to have teachers, supervisors, and administrators meet to discuss the results of the item analysis for specific groups of students and jointly decide on learning strategies to remedy the indicated areas of weakness.

Some caution must be exercised, however, when using an item analysis for the above indicated purpose. Scores for clusters containing relatively few items and scores for individual items do not have very high reliability. It is suggested therefore, that they be considered jointly with other available information such as: grades on homework and classwork; and teacher, supervisor, or administrator's observations of students' typical performance in the curriculum area under consideration.

Diagnosing learning difficulties for individual students is a somewhat different matter. Here the report that is provided by the publishers is a student response record which indicates for each student the items on which he/she scored correctly and incorrectly. Sometimes it also indicates which incorrect responses the student selected. As with the group item analysis caution should be exercised in using student response records for typical norm-referenced or criterion-referenced tests. These tests can provide a good indication of how well a student has mastered the total body of knowledge covered by the test. However, these tests simply do not provide enough items on each topic to make possible an indepth diagnosis of specific difficulties. These tests are excellent as survey instruments to locate possible areas of strengths and weaknesses. If one wants to pin-point specific difficulties, however, it is desirable to go to a more detailed test, one specifically designed for diagnosis. A number of very good diagnostic tests are available from the major publishers in the basic skills areas. If one wishes specifically to determine the exact kinds of difficulties a student is having, it is desirable to use a test of this type. A procedure that often proves to be quite useful is: first, use the student response record from a typical standardized achievement test to roughly determine which students are having difficulty; then, only for those students who appear to be having difficulty, administer a much more detailed diagnostic test. In most instances, it is not desirable to administer these diagnostic tests to all students because they are too long and time consuming. If a student is having serious difficulties of some sort, however, then it is advantageous to determine in detail what those specific difficulties are. In such a case the time involved in administering a diagnostic test is well spent. The main difference between the diagnostic test and the usual standardized achievement test is simply that the diagnostic test is longer and has more items to measure each specific skill. In that way the reliability of the measuring process for each sub-skill in the test is increased considerably.

Good examples of both the item analysis for groups and the individual student response record can be found in the results of the New Jersey Educational Assessment Program.

Detecting Discrepancies Between Potential And Achievement

The standardized tests that are most useful for detecting discrepancies between potential and achievement are tests of scholastic aptitude and tests of academic achievement. Some educators believe that scholastic aptitude tests measure innate potential whereas academic achievement tests measure the effects of the learning environment provided in the schools. This is not true. There is no way of measuring innate potential separate from the effects of prior learning. All psychological tests measure the individual's current behavior which undoubtably has been influenced by prior learning. However, tests of scholastic aptitude tend to be more general in nature than achievement tests and are used as a measure of one's ability to improve as a result of instruction. They are used to predict subsequent performance. Achievement tests are more specific in nature and tend to be used to evaluate an individual's status upon the completion of training. The emphasis in the case of the achievement test is on what the individual can do at the time. Generally scores on scholastic aptitude tests are used as predictors of scores on achievement tests.

How can we use the results of these two types of tests to get some indication of whether or not an individual is achieving his/her academic potential? First of all, we need to have the scores for these two types of tests expressed in some sort of common unit. If we have the result of a scholastic aptitude test expressed in the form of an I.Q. score and the results of the student's academic achievement test reported in national stanines, we might have some difficulty in determining exactly how the two are related. It would probably be useful to have both scores expressed in similar units. For example, both scores might be expressed in terms of national percentiles based on the same norms group; or both scores might be reflected in terms of stanines based on the same norms group.

Let us start with a simple example, suppose that the scores we have for scholastic aptitude and academic achievement are both reported in terms of local stanines. It would be unrealistic to expect that a student's scores on all of the subtests for curriculum areas of academic achievement would exactly equal his/her scholastic aptitude score even if he/she were achieving at a level commensurate with his/her potential. Fluctuations in an individual's score on any test are likely to occur as a result of chance factors irrelevant to the characteristics being measured. The **standard error of measurement** is a prediction of the range of fluctuation likely to occur in a single individual's score as a result of these irrelevant, chance factors. Although there may be a very high positive correlation between scores on an achievement test and scores on a scholastic aptitude test, the correlation between them will very likely be less than perfect. The **standard error of estimate** indicates the margin of error to be expected in predicting one score from another when the correlation between the two sets of scores is less than perfect. When attempting to predict achievement test scores from scholastic aptitude scores one must take into consideration the various sources of random error: the **standard error of measurement** for each of the instruments plus the **standard error of estimate** for predicting one score from the other. The **standard error of measurement** for the two instruments may be found in the manual available from the publisher and the **standard error of estimate** for predicting one set of scores from the other may be computed.

Despite these limitations, one might still want to examine the relationship between a student's achievement test scores and scholastic aptitude scores as **one possible** indicator of whether or not the student is achieving on a level commensurate with his/her academic potential. As is the case in making any inferences from test data, the test data should not be considered as the sole criterion but merely as adjunct information. It is recommended that it be considered together with other data such as teacher judgement based on first-hand knowledge of the student's typical daily performance.

Assuming reasonably small standard errors of measurement on both the aptitude and achievement tests and a very high positive correlation between them, it might be reasonable to expect that a student's stanine score for academic achievement would tend to fall within plus or minus one stanine of his/her score for scholastic aptitude. For example, if his/her stanine score for the scholastic aptitude test was six we would expect that his/her scores in the area of academic achievement on the various subtests would most likley fall in the range of five through seven. If a score fell below five, we might have reason to suspect that the student is achieving below his/her potential. If for some reason a score fell above seven we might think that the student was perhaps

making some extraordinary effort in that area and therefore achieving more than we would normally expect him/her to achieve.

A simple chart constructed to show standardized achievement scores on the various subtests and how they relate to scholastic aptitude is shown in Figure 2. The two heavy horizontal lines reflect the boundaries in which we would expect most achievement test scores to lie. Anything above the upper boundary would be an indication of exceptional achievement in relation to potential. Anything below the bottom horizontal line would be an indication of possibly achieving below one's potential. In this example, the student has a stanine score of five for scholastic aptitude.

Similar kinds of things could be done with other types of scores. If we were using national percentiles, we might have some sort of graphic arrangement of percentile scores. We might decide that we would expect the student's national percentile score for academic achievement in each curriculum area to fall within plus or minus 10 percentile points of his/her percentile score for scholastic aptitude. Some testing companies will provide this type of service if you use their tests for both scholastic aptitude and academic achievement. They will report to you whether or not a student's achievement scores reflect achievement commensurate with his/her potential based on his/her age, grade level, and scholastic aptitude scores. The multiple regression techniques used for this purpose statistically allow for the previously cited sources of random error.

FIGURE 2
Standardized Test Profile (Grades 4-6)

Sample City Public Schools
Sample City, Any State

In an attempt to share as much information as possible regarding pupil progress, the form below has been devised to give parents a graphic picture of the relationship of a pupil's ability to his/her achievement. Although it is the intent of the school to share and discuss this information in parent conferences, it should be understood that the scores of all standardized group tests are such they they may not be valid in individual cases. Therefore, information concerning a pupil's classroom performance should be weighed in the same balance as the test profile.

Student *Mary Somebody*

Grade *4* **School** *Sample Elementary School*

Year *1977-1978* **Teacher** *Ms./Mr. Anybody*

	Stanine	Word Meaning	Paragraph Meaning	Spelling	Word Study Skills	Language	Arithmetic Computation	Arithmetic Concepts	Arithmetic Applications	Social Studies	Science	Stanine	
Above Average	9											9	Above Average
	8											8	
	7											7	
Average	6											6	Average
	5	X	X		X			X	X	X	X	5	
	4			X		X	X					4	
Below Average	3											3	Below Average
	2											2	
	1											1	

Assessing Growth In Academic Achievement

Standardized test results can provide considerable help in assessing student growth in academic achievement by answering two very important questions. The first is: What new skills has a student mastered and what new knowledge has he/she gained over a specific period of time? The second is concerned with how this growth compares with that of other students. The first of the two questions can best be answered by examining the results of achievement tests of the criterion-referenced type. As mentioned previously, criterion-referenced tests are concerned with comparing a student's mastery of specific skills or a specific body of knowledge with some predetermined standard. They do not attempt to compare a student's performance with that of other students.

In order to assess growth in academic achievement using a test of this type, results of tests administered at two different points in time are needed. By comparing a student's performance on the first test with his/her performance on the second test, one can determine what new skills, new knowledge, or new objectives, this student has mastered during the time interval between the two tests. A student response record is a report that shows which items a student answered correctly on a test. By examining the student response record for both of the tests one can see specifically which items the student answered correctly the second time that he/she did not answer correctly the first time.

If the results are reported in terms of clusters of items with each cluster measuring a particular skill or a particular related body of knowledge, an assessment can be made concerning what progress, if any, has been made in each of these specific areas. By looking at the percentage of correct responses on the total test for both the first test and the second test, one can also see if any change or growth has occured for the overall body of knowledge covered by the total test. Not only can this be done for individual students, but by examining average results it can also be done for groups of students, whether it be a class, an entire grade level within a school, or the entire school district. For the group in question, the percentage of students who scored correctly on each item during the first test can be examined and compared with the percentage of students who scored correctly on each item during the second test. In a similar manner, the average percent of correct responses on the total test for the first and second tests can be compared. In this way we can assess how well the group as a whole is progressing in academic achievement in the curriculum area measured by the test.

When assessing growth by comparing scores on a pretest administered at the beginning of the academic year and a post-test administered at the end of the year, some caution is advised. If the same form of the test is used both pre and post, scores on the post-test may be somewhat inflated due to a memory or practice effect. Students may respond correctly to some items on the post-test simply because they have responded to them before and remember the correct response. To eliminate such practice and memory effects it is desirable to use two **different** parallel forms of the same test. The parallel forms have different test items but are of similar length and difficulty and measure the same content and skills.

Results from criterion-referenced tests can be quite useful in reporting individual student progress to parents. These results make possible very detailed reporting of what specific skills and knowledge the student has or has not mastered. In a like manner, group results can be reported to Boards of Education or other interested community groups using group averages. To answer the second question concerning how this growth in academic achievement compares with that of other students, the results of norm-referenced tests can be used. These tests are specifically designed to compare a student's progress with that of other students.

For example, if we are using national percentile scores and looking at the results of a test that was administered at the end of fourth grade to a particular student and another similar achievement test that was administered to the same student at the end of the fifth grade, progress can be assessed in the following manner. If a student scored at the national 50th percentile at the end of fourth grade, this would indicate that his/her achievement was average in the norms group for students who took this test at the end of the fourth grade. If he/she scored at or close to the 50th percentile again, one year later, on his/her fifth grade test one might conclude that his/her growth during that year from the end of fourth grade to the end of fifth grade was just about average in comparison to the students in the norms group. If however, the student scored at the 65th percentile at the end of fourth grade and only scored at

the 50th percentile at the end of fifth grade, he/she probably made less than average growth during that period of time. His/her position in the norms group was somewhat lower the second year.

If stanine scores were used and the student received a stanine score of 4 at the end of fourth grade and again a stanine score of 4 at the end of fifth grade, he/she would be maintaining the same position in the norms group both years indicating average growth during that period of time. If however, during the one year period he/she went from a stanine score of 4 to a stanine score of 6, he/she probably experienced greater than average growth during that period of time. If he/she went from a stanine score of 6 during the first year to a stanine score of 4 the second, his/her position in the norms group has dropped somewhat and he/she has probably experienced less than average growth in academic achievement during that year.

The same sort of assessment may be made for groups of students. However, when using norm-referenced tests it might be desirable to use more information than average or mean scores. Three points in the distribution of scores might be examined: the 25th percentile, the point at which a quarter of the students score below and three-quarters score above; the 50th percentile, the point at which half of the students score below and half score above; and the 75th percentile, the point at which three-quarters of the students score below and one-quarter score above. This provides a means for examining the achievement levels of the lower, average, and higher performing students, respectively. A useful technique is to find the raw score in the local distribution for each of these percentile points (25th, 50th, and 75th) and then see where that particular raw score falls in the distribution of scores for the national norms groups.

For example, suppose the particular raw score that was at the 50th percentile in the local school district turned out to be at the 60th percentile nationally. This would indicate that the students in this particular school district are performing somewhat higher than the national norms level, because in that particular school district 50 percent of the students scored at or above that point, whereas in the nation only 40 percent of the students scored at or above that point. Whenever the national percentile for a given raw score is above that of the corresponding local percentile (25th, 50th, or 75th) the local school district is performing better than the national norms group. When the national percentile for a given raw score is below that of the corresponding local percentile, the local school district is performing below national norms. Where the two percentile points, local and national, are equal, the local district is performing exactly at the national norms level. Figure 3 is an illustration of a device that uses this procedure. However, in Figure 3 the results are given for three years. By looking at the results over a period of three years, one can begin to determine whether there is an upward or downward trend, whether growth is greater than average at each of the quartile points or whether it is less than average.

Grouping Students For Instruction

Students may be grouped for instruction in a great variety of ways. They may be grouped homogeneously in regard to some specific attributes such as academic ability, reading achievement, or mathematics achievement. Thus, any given class will contain a narrow range of differences in regard to the attribute in question. They may be grouped heterogeneously as to some specific attribute, so that any given class contains a wide range of differences in the attribute in question. They may simply be grouped randomly or in alphabetical order. Here, however, we are concerned with systematic grouping according to some attribute that may be measured by standardized test results.

Before we proceed further, a word of caution should be added. Rarely, if ever, should standardized test results be the sole basis for any form of grouping. Other sources of data such as cummulative records, report card grades, and teacher judgement should also be used.

The procedure that follows is often a useful one in grouping students for instruction. Take the scores on the attribute in question and arrange them in rank order from high to low for the given group of students. Then look at the ranking of students in terms of other data such as teacher judgement, report card grades, test results from previous years, and so forth. Based on the additional data, make any rearrangements that you deem necessary in the ranking. After all the data has been considered and students have been rearranged in

FIGURE 3

Brand X
Achievement Tests: National Percentiles
Grades 3 - 8 Administered in
April, 1977

Sample City School District

Sample City School District	Local Percentile	Vocabulary Percentile Year			Reading Compre. Percentile Year			Word Study Skills Percentile Year			Math Concepts Percentile Year			Math Comput. Percentile Year			Math Applic. Percentile Year			Spelling Percentile Year			Language Percentile Year		
		'75	'76	'77	'75	'76	'77	'75	'76	'77	'75	'76	'77	'75	'76	'77	'75	'76	'77	75	'76	'77	'75	'76	'77
Grade 3	75%ile	84	84	88	86	86	89	89	89	89	88	90	90	88	92	92	84	90	90	84	88	92	84	84	90
	50%ile	70	70	77	68	70	76	72	82	82	72	76	76	64	76	76	68	72	78	66	70	74	68	68	74
	25%ile	50	54	58	48	50	54	52	58	62	42	52	58	42	50	56	46	46	50	40	40	46	46	44	52
Grade 4	75%ile	84	84	86	85	82	86	84	86	86	82	82	86	84	82	86	83	80	84	84	84	88	81	84	86
	50%ile	65	62	70	65	60	70	63	62	72	63	60	70	63	62	74	62	58	66	66	66	74	62	62	70
	25%ile	44	44	50	42	40	46	35	34	40	39	40	46	37	38	50	41	36	44	43	42	50	38	40	44
Grade 5	75%ile	79	80	80	86	86	88	84	88	--	83	88	88	80	86	88	81	86	82	82	88	88	86	88	89
	50%ile	61	62	66	67	70	70	66	72	--	59	66	70	60	64	72	57	66	62	65	68	72	67	70	74
	25%ile	42	44	48	43	44	50	44	48	--	37	40	56	35	40	50	38	38	38	42	40	46	41	44	50
Grade 6	75%ile	86	82	80	87	84	88	86	86	82	86	90	92	80	82	89	80	80	80	87	84	88	87	84	88
	50%ile	64	62	58	65	66	72	70	70	50	67	72	77	59	66	70	61	62	62	65	64	66	67	66	70
	25%ile	42	38	32	44	42	46	49	50	24	42	42	52	37	44	48	39	38	38	42	42	42	44	42	46
Grade 7	75%ile	82	84	84	80	84	82	--	--	--	85	89	88	79	84	84	81	84	86	81	83	89	84	85	86
	50%ile	64	66	66	62	64	64	--	--	--	65	70	66	57	64	66	62	66	66	56	61	60	63	68	68
	25%ile	44	48	46	41	44	42	--	--	--	38	48	48	31	36	40	39	40	40	30	36	36	32	42	40
Grade 8	75%ile	84	84	88	80	83	82	--	--	--	85	86	86	81	81	82	83	84	86	83	83	84	83	83	86
	50%ile	64	68	72	60	64	64	--	--	--	66	68	68	58	58	62	65	64	68	59	58	60	63	64	68
	25%ile	43	48	46	37	38	44	--	--	--	47	44	44	36	32	36	42	38	44	34	30	36	36	38	46

the best rank order on the attribute in question, form the groups based on this particular ranking, whether they be deliberate homogeneous groups or deliberate heterogeneous groups. The most heavily weighted factor in this process should probably be teacher judgement since teachers have first-hand knowledge of students typical daily performance. Test scores simply offer a convenient starting point by providing a criterion easily ranked from high to low.

If, for example, we were going to arrange a given group of about 90 students into three classes of 30 students each in a homogeneous manner based on the attribute in question, we would simply take the top 30 and put them in one group, the next 30 and put them in the second group, and the bottom 30 and place them in the third group. If we were to deliberately arrange the students in heterogeneous groups with a wide range of differences in each, we might use a procedure like the following. Suppose again we have 90 students and we wish to arrange them in three classes of 30 each, so that they are heterogeneously grouped this time according to the attribute in question. We might divide this group of 90 students into three sections, a top, middle, and bottom section. In forming our three classes, we might then randomly take 10 students from each of these three sections to form each class. In that way we would insure a wide range of differences on the particular attribute in question.

Some specific examples might help to clarify this concept. Let's suppose we have a school staff that believes that students can best learn in an atmosphere where students of a variety of ability levels interact together. They might feel it would be desirable to group students heterogeneously according to academic ability as measured by I.Q. Let us again suppose that we have three classes at a given grade level and that we wish to group these students into three heterogeneous classes based on I.Q. scores and other data. Our first step might be to arrange the students in rank order from high to low based on their I.Q. scores. Teachers might then scrutinize this ranking from the viewpoint of students' classroom performance and they might want to re-rank some students based on their own judgement about the student's ability. Then when a final ranking was achieved, the classes might be formed as follows: the ranking might be divided into three categories, a high category, a middle category, and a low category. When classes were selected, one-third of each class might come from each of these three categories, thus insuring a wide range of academic ability. The classes when formed might also be based on other characteristics such as compatibility of student personalities, or compatibility of student personality with teacher personality, or an infinite variety of other factors.

Since students are often grouped homogeneously for instruction in the basic skills, let's take an example involving reading ability or reading achievement as measured by standardized achievement test scores and teacher judgement. Here we might take the given group of students and arrange them in rank order from high to low based on their achievement scores in reading, a composite score in reading, involving reading comprehensive, word study skills, and vocabulary. After this ranking had been achieved, teachers might want to do some rearranging based on their knowledge of the students' typical classroom performance in reading. After the final ranking had been achieved from high to low considering all of the various inputs (teacher judgement, report card grades, standardized test results, etc.), then classes might be formed based on this ranking. If we were going to form three classes for a particular grade level, we would simply take the top third for one class, the second third for the next class, and the bottom third for the final class.

Using procedures of the type indicated, standardized test results can often be an excellent starting point for forming instructional groups. Again, however, the reader is cautioned that test results should not be the sole basis for grouping. Other valuable sources of data should also be considered.

Planning Instructional Activities

During the teaching-learning process considerable time and effort is sometimes wasted. Teachers sometimes attempt to teach students topics they are not ready to learn due to lack of prerequisite background or other factors. Sometimes teachers unknowingly attempt to teach students things that they already know. Much of this waste of time and effort can be avoided by careful instructional planning which makes use of the information provided by the results of criterion-referenced tests.

Suppose that early in the academic year we are planning an instructional program for a given group

of students in one of the major curriculum areas. Let us further suppose that we have available to us a criterion-referenced test which has sufficient items to measure all of the important instructional objectives that will be covered during the year in that particular curriculum area. Our first step in the planning process might be to administer the criterion-referenced test and examine the results, particularly the results of the item analysis. The item analysis will tell us the percentage of students who scored correctly on each of the test items. If our groups of items are arranged according to the instructional objectives they measure, they will further tell us which instructional objectives have already been met by the majority of students and which have not.

We will probably find that there is some overlap in the instructional objectives that are specified for the grade level we are dealing with, and those specified for the previous grade level. In examining our test results, then, we will probably find there are some objectives that have already been met by a substantial proportion of the students in our group. We will probably further find that there are a large number of objectives that have not been met by the majority of students in our particular group. We will probably find also a number of objectives that have not been met by the majority of students, but have been met by a significant proportion of them. As we examine our test results, we might want to classify our objectives into each of the three categories just mentioned: (1) identify those objectives which have been met by the great majority of students in the class; (2) identify those objectives which have not been met by any significant proportion of students in the class; and (3) identify those which have been met by a significant proportion of students within the class but certainly not by the majority.

With this information we can begin to form our instructional program. We can plan some brief review for those objectives that have been met by the majority of students, and heavy instructional units for those objectives that have not been met by any significant proportion of students. For those objectives that have been partially met by the group, we may form some sort of grouping process in which we sub-group according to individual needs in the areas related to these objectives. By using the individual student response records which indicate which specific items each student answered correctly on the test, we can begin to group students for instruction on the objectives which have been met by a significant proportion of students within the class. Then we can begin to form instructional groups in those particular areas.

We may find a small number of students within the class whose scores are particularly low on the criterion-referenced test. For these students it may be necessary to go back to the student response record of the criterion-referenced test of the previous grade level. This will enable us to determine whether or not they have the prerequisite background to deal with the instructional activities that will be used for the great majority of students within the class. If they are found to be lacking in prerequisite background, it may be necessary to do some re-teaching of the objectives that are designated for the previous grade level and other grade levels below that.

As the instructional program proceeds during the year, it may be desirable to supplement the information provided by the initial criterion-referenced test with that of teacher-made tests which are given prior to and at the conclusion of each unit of work. Each unit of work will probably consist of one or two objectives which are somewhat related, or perhaps of a cluster of related objectives. The pre-test in each case will give the teacher information on how ready students are to proceed with the unit of work and the post-test will provide some indication at the conclusion of the unit as to whether or not students have indeed mastered the material presented. If the pre-test shows that a substantial number of students do not have the readiness to proceed with the unit of work in question, it may be necessary to do some re-teaching of the objectives of the previous grade level which are prerequisites for this particular unit. If, on the other hand, the pre-test shows that the youngsters have already mastered a substantial amount of the content of the unit in question, then it may be necessary to move ahead more rapidly than originally anticipated; and perhaps, for the purpose of review, touch lightly on those objectives that have been mastered placing major emphasis on those objectives which have not already been mastered. In some instances the pre-test results may be so good that it is necessary to simply go to the post-test right away and see whether indeed, the students have already mastered the content of the unit in question.

At the end of the year it may be desirable to administer again the same criterion-referenced test that was administered at the beginning of the year or perhaps an alternate form. This may provide a means for assessing how effective the instructional program has been for the group of students in

question during the academic year. By reviewing the results for the test administered at the beginning of the academic year and comparing it with those for the test given at the end of the academic year, it should be possible to assess which instructional objectives have been met and which have not. This should provide considerable information concerning which instructional materials and strategies have proven to be effective and should provide information for revision of the instructional program for use with subsequent groups of students.

When used in the manner described, the results of criterion-referenced tests can provide important information for the process of planning instructional activities.

Determining Reasonable Achievement Levels

Norm-referenced and criterion-referenced achievement tests can both provide useful information for determining reasonable levels of achievement. Since many of the concepts involved in using norm-referenced tests for this purpose were covered in one of the preceding sections, this section will concern itself primarily with the use of criterion-referenced tests.

In using the results of criterion-referenced tests for this purpose, the general question we will concern ourselves with is: What specific body of knowledge and what specific skills is it reasonable to expect a given group of students to master at a given grade level? It would be very convenient if some simple formula or guideline existed for this purpose. However, this is not the case. Unfortunately there are no easy answers and no widely accepted procedures for use in answering this question.

The logical place to begin, however, is with the situation as it presently exists. For the particular school or school district in question, a criterion-referenced test might be selected on the basis of high content validity for the current curriculum. It might then be administered to the given group of students at the end of the academic year to see just what level of mastery currently exists. Prior to examining the item analysis for the test results, it might be useful to have teachers indicate the percentage of students they expect to score correctly on each test item. These expectations can then be compared with the actual results to see where discrepancies exist. Next, it might be useful to use opinions of supervisors, administrators, teachers, and students to identify some specific classes within the school or school district where the level of instruction seems to be exceptionally good. The test results for these specific sections or classes can then be compared with the general average results for the district to see whether any differences actually exist between them. If the test results for these classes identified as exemplary do indeed show much greater mastery for the educational objectives in question, they might serve as some sort of ideal guideline for district; something toward which the district can strive.

The process of determining reasonable achievement levels can never be static. Each year it is necessary to again examine the results of the criterion-referenced tests and critically examine the objectives, the instructional materials being used, and the instructional strategies, to determine what changes might need to be made. In some instances no matter how effective the instructional techniques and how exemplary the instructional materials, certain objectives still cannot be met for a given group of students at a given grade level. In that case, it is necessary to modify the curriculum and move the educational objectives in question to a higher grade level. Also as curriculum changes are made, corresponding changes must be made in the testing program to maintain content validity.

About the Author...

Eugene J. Travers is Supervisor of Research and Development for the East Brunswick Public Schools, East Brunswick, New Jersey. He received B.Mus. and M.Mus degrees from Westminister Choir College and M.Ed. and Ed.D. degrees from Rutgers University.

His current areas of specialization are educational statistics and measurement, and research and evaluation design. Prior to his present position he served as an elementary teacher and Coordinator of Mathematics and Science for the East Brunswick Public Schools and as an evaluation intern for the New Jersey Department of Education.

He has taught courses in mathematics, education, and psychological statistics for Kean College of New Jersey and a course in psychological measurement for University College of Rutgers University. He also has served as a consultant in mathematics education for Science Research Associates, Inc.

PUBLICATIONS IN THIS SERIES INCLUDE:

Volumn I.	Purpose of Testing	*David W. Pearsall*
Volumn II.	Test Selection and Evaluation of Tests	*Malcolm J. Conway*
Volumn III.	Organizing and Managing a District Testing Program	*Robert L. Jackson, II*
Volumn IV.	Some Uses of Standardized Test Results	*Eugene J. Travers*
Volumn V.	Motivation and Test Wiseness	*David F. Englehardt*

Published by the New Jersey Department of Education, 225 West State Street, Trenton, New Jersey 08625. The opinions expressed in this publicaition are those of the invited authors and do not necessarily reflect the position or policy of the New Jersey Department of Education.

ACKNOWLEDGMENTS

Thanks are extended to the Planning Committee for this publication and to the editorial review panel consisting of Dr. Curtis Banks, Princeton University; Dr. Douglas Penfield, Rutgers University; and, Dr. Paul Van R. Miller, Lehigh University.

Appreciation is also expressed to Ms. Vivian Sobke for preparation of the manuscripts.